Marginalized Latent Semantic Encoder for Zero-Shot Learning

Zhengming Ding[†] and Hongfu Liu[‡]
[†]Department of CIT, Indiana University-Purdue University Indianapolis, IN, USA
[‡]Michtom School of Computer Science, Brandeis University, MA, USA
zd2@iu.edu, hongfuliu@brandeis.edu

Abstract

Zero-shot learning has been well explored to precisely identify new unobserved classes through a visual-semantic function obtained from the existing objects. However, there exist two challenging obstacles: one is that the human-annotated semantics are insufficient to fully describe the visual samples; the other is the domain shift across existing and new classes. In this paper, we attempt to exploit the intrinsic relationship in the semantic manifold when given semantics are not enough to describe the visual objects, and enhance the generalization ability of the visual-semantic function with marginalized strategy. Specifically, we design a Marginalized Latent Semantic Encoder (MLSE), which is learned on the augmented seen visual features and the latent semantic representation. Meanwhile, latent semantics are discovered under an adaptive graph reconstruction scheme based on the provided semantics. Consequently, our proposed algorithm could enrich visual characteristics from seen classes, and well generalize to unobserved classes. Experimental results on zero-shot benchmarks demonstrate that the proposed model delivers superior performance over the state-of-the-art zero-shot learning approaches.

1. Introduction

Visual data analytic has achieved tremendous improvements recently, as the rapid explosion of data scales and continuously-improved learning models. Traditional visual recognition systems almost pursue the supervised strategies, that require a great number of well-annotated instances to seek a high-performance model. Unfortunately, it is expensive and even prohibitive to collect enough training samples for an effective model, especially when these samples need fine-grained annotations. Hence, it is appealing and essential to build such recognition systems that can identify novel categories in the test stage with limited or even no instances accessible in the training process.

Zero-shot learning (ZSL) has been surging recently, which catches great attention for its promising performance

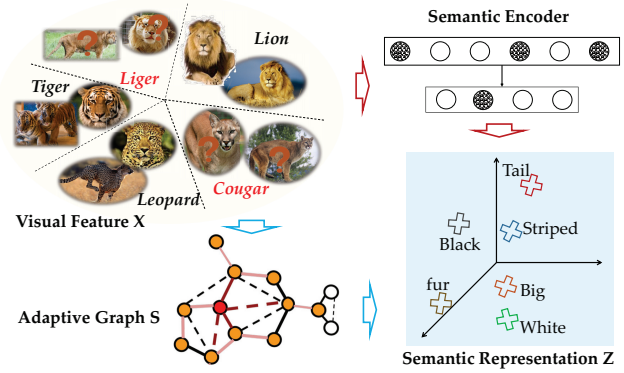


Figure 1. Illustration of our marginalized latent semantic encoder, in which a semantic encoder is built to bridge visual features in marginalized corruption $\mathbb{E}(\tilde{X})$ and the latent semantics Z with $W\mathbb{E}(\tilde{X}) \approx Z$. Furthermore, latent semantics are learned over the given semantics A through an adaptive graph ($Z \approx AS$).

in generalizing knowledge from observed objects to unseen objects [22, 8, 33, 11, 6, 17, 14, 31, 3, 26, 7, 28]. In fact, ZSL is motivated by the human cognitive learning mechanism in identifying unknown classes. ZSL attempts to discover the intrinsic visual-semantic mapping from observed objects and generalizes it for unobserved categories. One of the most frequently-adopted approaches is to embed visual features and their corresponding semantics of seen classes into the same common space to couple the semantic gap across two, which expects that the unseen classes with their semantics and visual samples are also embedded in the same space. Most existing ZSL models focus on seeking the visual-semantic function only relying on the provided visual data and its semantics [4, 22, 12, 13]. The visual-semantic function could simply be a linear mapping [23], or dual linear mappings [4], or even complex non-linear functions, including dictionary learning [8], auto-encoder [12, 17, 3, 34], and generative models [38, 31, 9, 7], where generative models are usually promising to augment the space of seen classes and more likely to cover that of unseen classes in the training stage.

Although the existing ZSL methods achieve some promising results in generalizing the seen knowledge to un-

seen ones [12, 17, 3, 34], there still remains two degenerating points. First of all, it presents a general challenge from no training data for the unseen classes, which leads to the difficulties for model selection. The domain shift across seen and unseen classes would prevent the generalization ability of the learned visual-semantic function. Thus, how to learn an effective and compatible visual-semantic mapping on the observed objects is the key problem in ZSL problem. Secondly, the information based on observed samples is not always sufficient to learn the visual-semantic mapping. On one hand, the semantic attributes are subjective to be annotated and not enough to span the visual feature space; on the other hand, the visual-semantic mapping is learned only on the seen categories, where the different visual distribution on seen and unseen categories obstacles the effective generalization in the test stage. To this end, tremendous efforts have been taken to handle the above challenges[38, 31, 9, 7]; however, most of them ignore the huge potential in the latent semantic representation for a more generic visual-semantic mapping learning.

In this paper, we develop a novel Marginalized Latent Semantic Encoder (MLSE) to deal with the previously-mentioned two zero-shot obstacles (Figure 1). Our main assumption is that the latent semantic representation could better describe the visual samples compared with human-annotated ones, and generic semantic encoder is able to better capture the unseen knowledge by augmenting visual space of the seen classes through marginalized denoising strategy. Moreover, we exploit a sparse residual constraint to purse a meaningful semantic embedding space and guide the latent semantic representation learning. To sum up, we highlight our contributions as:

- First of all, we derive a generic encoder to adapt the intrinsic knowledge and shared features from the observed classes under a marginalized augmentation. Therefore, a generic semantic encoder could cover more knowledge for the unseen categories, and thus generalize well in the test stage.
- Second, we automatically learn new latent semantics to seek more efficient prototypes from known classes through an adaptive graph reconstruction strategy over given semantics. Hence, our model is able to learn more effective information with the given human-annotated semantics.
- Finally, we further adopt a sparse reguralizer to constrain the adaptive graph learning with preserving the original intrinsic information and removing the outliers and noising factors. Therefore, our model is able to effectively learn the latent semantics.

2. Related Work

Zero-shot learning (ZSL) targets at learning models of visual concepts with no evaluation data of the concepts. As visual knowledge from such unknown evaluation classes is inaccessible in the training process, ZSL needs external semantics to compensate for the unknown visual information. So far, attribute-based descriptions are widely used to define the shared characteristics across various categories [20, 21], which is an intermediate domain to link the visual features with their semantics.

Early ZSL explores the attributes within a two-stage approach to predict the label of a given image from the unseen classes. Generally speaking, the attributes of any given image are assigned in the first stage, then its class label is inferred by searching the class-attribute table using the nearest neighbor classifier. Direct Attribute Prediction (DAP) and Indirect attribute prediction (IAP) are two pioneering studies, which adopt the hidden layer of attributes as variables decoupling the images from the layer of labels [15]. However, such two-stage approaches suffer from distribution difference between the intermediate and target task, since target task is to assign the class label while intermediate task would consider to obtain attribute classifiers.

Recent advances of ZSL seek a direct mapping from a visual feature space to a semantic space. Along this line, Akata et al. optimized the structural SVM loss to achieve the bilinear compatibility [2]. Furthermore, they proposed to build a bilinear compatibility function across the visual and the semantics via a ranking loss [1]. On the other hand, Romera-Paredes et al. exploited the square loss to obtain the bilinear compatibility and explicitly regularizes the objective [23]. Recently, Jiang et al. also employed a dictionary learning framework to seek the latent attributes, which was not only discriminative but also semantic-preserving [11]. Liu et al. explored a semantic auto-encoder with rank constraint on the projection matrix to preserve more intrinsic structure [17]. Some generative models are proposed by seeking a generator as the visual-semantic mapping function [38, 31, 9]. They mainly explore the conditioned generator on semantics to synthesize more visual features for seen classes, and thus they have a better chance to mitigate the domain shift in visual space between seen and unseen classes. However, generative models are usually hard to train due to its min-max optimization.

Moreover, another ZSL direction is to embed both the visual and semantic features into a shared intermediate space. Following this, Zhang et al. mapped visual features and semantic features into two different latent spaces, and measured their similarity through seeking one bilinear compatibility function [36]. Besides, Changpinyo et al. explored a hybrid model and constructed the classifiers of unseen classes by taking the linear combinations of base classifiers, which are trained in a discriminative learning framework from seen classes[5].

Unfortunately, most existing ZSL approaches pay less attention to discriminative information for the unknown cat-

egories considering the high within-class variability, and therefore, they would fail to uncover the common semantics cross seen and unseen classes. Differently, we assume the provided semantics are not enough to describe the visual objects and thus aim to seek a better latent semantic representation. Simultaneously, we learn a generic semantic encoder with marginalized augmentation strategy to effectively handle the domain shift and discover the shared discriminative features across the seen and unseen categories.

3. The Proposed Algorithm

In this part, we discuss our novel marginalized semantic encoder with latent semantic representation for effective zero-shot learning.

3.1. Preliminaries & Motivation

Considering there are C seen categories with n labeled instances $\mathcal{D} = \{X, A, y\}$ and C_u unseen categories with n_u unlabeled instances $\mathcal{D}_u = \{X_u, A_u, y_u\}$. Each instance is represented with a d-dimensional visual feature vector. $y \in \mathbb{R}^n$ and $y_u \in \mathbb{R}^{n_u}$ denote class labels for the seen and unseen categories, respectively. More specifically, the seen and unseen categories are non-overlapped in term of category information, that is, $y \cap y_u = \emptyset$. Thus, semantic representations make up for this challenge, where $A \in \mathbb{R}^{m\times n}$ and $A_u \in \mathbb{R}^{m\times n_u}$ are the m-dimensional semantics for seen and unseen categories, respectively. For the seen categories, A is given for visual feature X, which is labeled by either binary or continuous attributes representing its corresponding class label y. By comparison, A_u has to be predicted as the unseen categories are not annotated. The intuition of ZSL is to learn a visual-semantic function to discover the relationship across the visual features and the individual dimensions of the semantic features. Due to the distribution divergence across seen and unseen categories, it is essential to mitigate this challenge during the visual-semantic function learning.

Since seen categories X and unseen categories X_u are sampled from various visual feature spaces; fortunately, A and A_u compensate by sharing some common semantics with each other. Take attribute-based semantics as an example, both seen and unseen categories can be described with human-annotated attributes in various values either binary or continuous. Besides, we notice the human-provided semantics are not sufficient to comprehensively describe the visual samples. To this end, we propose our marginalized latent semantic encoder to handle these two challenges. First, we explore to diversify the feature space of seen classes during model training by using the marginalized desnoising strategy. Second, latent semantic representation is sought to better describe the visual samples jointly with an adaptive graph learning.

3.2. Generic Semantic Encoder Learning

A natural way to enhance the generalization of a visual-semantic function using a corrupting distribution is to explore the spirit of [18] by selecting each element of the training samples and corrupting it k times. For seen visual features X, this results in corresponding corrupted observations $\tilde{X}_l$ (with $l = 1, \cdots, k$). Thus, we propose a semantic encoder to encode each corrupted $\tilde{X}_l$ with semantic representation A as follows:

$$\min_{W} \frac{1}{k} \sum_{l=1}^{k} \|W\tilde{X}_l - A\|_{\mathrm{F}}^2, \ \text{s.t.} \ W^\top W = \mathrm{I}_m, \tag{1}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm and $\tilde{X}_i$ is the l-th corrupted version of X. Note that the orthogonal constraint $W^\top W = \mathrm{I}_{\mathrm{m}}$ ($\mathrm{I}_{\mathrm{m}} \in \mathbb{R}^{m\times m}$) is imposed to avoid trivial solutions.

Although attribute semantics are widely-used in the classification problem, two issues need to be taken into account. First of all, the human-annotated attributes do not always achieve the similar importance for discrimination, hence, it would be not desirable to seek more enriched semantics. Secondly, there are correlations among different attributes; hence, it is improper to learn every attribute individually. In other words, it is too strong to enforce A to be semantics output. Thus, we explore to learn new latent semantics to relax the constraint. Furthermore, we introduce the marginalized denoising strategy to consider the limiting case when k tends to be ∞. To this end, we explore the weak law of large numbers and reformulate $\frac{1}{k}\sum_{l=1}^{k}\|W\tilde{X}_l - A\|_{\mathrm{F}}^2$ to its expectation formula:

$$\begin{aligned} \min_{W,Z} \mathbb{E}\Big(\|W\tilde{X} - Z\|_{\mathrm{F}}^2\Big) + \alpha\|Z - A\|_{\mathrm{F}}^2, \\ \text{s.t.} \ W^\top W = \mathrm{I}_m, \end{aligned} \tag{2}$$

where $\mathbb{E}(\cdot)$ is the expectation operator and α is the trade-off parameter. The second constraint enforces Z to be similar to the given semantics A, which could help ensure that the learned semantic encoder depicts visual-semantic relation. The exception loss minimization results in data augmentation in learning the semantic encoder, which would improve the generalization ability of the proposed model, especially when dealing with zero-shot learning problem.

3.3. Adaptive Graph Guided Latent Semantics

Considering the phenomenon that samples from one category are lying over a complicated manifold, e.g., crescent manifold, it is clearly not a proper way to directly use its center as its prototype or exemplar to that class. Moreover, human-annotated semantics are usually not enough to comprehensively describe the visual samples. However, semantics across different categories should be shared. So, we adopt the manifold learning idea to uncover the latent

semantic representation based on a new reconstruction of provided semantics constrained a graph manifold. Thus, semantics of some instances lying in the same local manifold could compensate with each other in latent semantics learning.

To uncover more semantic knowledge, we propose an adaptive graph reconstruction term to align the latent semantics Z and the given semantics A as follows:

$$\begin{aligned} &\min_{W,Z,S} \mathbb{E}\Big(\|W\tilde{X} - Z\|_{\mathrm{F}}^2\Big) + \alpha\|Z - AS\|_{\mathrm{F}}^2 \\ &\text{s.t. } W^\top W = \mathrm{I}_m,\ \ \mathbf{1}_n^\top S = \mathbf{1}_n^\top,\ \ S \geq 0, \end{aligned} \tag{3}$$

where two constraints, i.e., $\mathbf{1}_n^\top S = \mathbf{1}_n^\top$ ($\mathbf{1}_n$ is an all-one n-dim vector) and $S \geq 0$, are used to guarantee the validity of the obtained graph coefficients.

To explore more intrinsic structure on S, we consider a residual minimization with pre-defined graph weight matrix H, which is calculated from a spectral dual-graph. To be specific, we explore the data structures from both visual features X and its corresponding semantics S. We build two k-nn graphs G_x and G_a based on visual and semantic features, respectively. We first use cosine similarity to calculate the weights of two graphs, i.e., H_x and H_a, then exploit a simple fusion strategy to achieve weight matrix H as $H = \frac{H_x + H_a}{2}$. However, the learned graphs G_x and G_a may suffer from arbitrary noise from the data. In the worst case, it would significantly affect the learning of latent semantic representation and further the semantic encoder. To this end, to promote structure information and suppress effects of noise data points, we first explore l_1-norm to constrain the residual between H and S in order to figure out the small number of abnormal weights caused by outliers or noisy samples. We expect most elements of S to be similar to H to preserve the original intrinsic structure while some to be different for outliers. Thus, we achieve a robust graph guided semantic encoder as follows:

$$\begin{aligned} &\min_{W,Z,S} \mathbb{E}\Big(\|W\tilde{X} - Z\|_{\mathrm{F}}^2\Big) + \alpha\|Z - AS\|_{\mathrm{F}}^2 + \beta\|S - H\|_1 \\ &\text{s.t. } W^\top W = \mathrm{I}_m,\ \ \mathbf{1}_n^\top S = \mathbf{1}_n^\top,\ \ S \geq 0, \end{aligned} \tag{4}$$

where β is the trade-off among three components. $\|\cdot\|_1$ is the l_1 operator of matrix to detect the outliers in the original dual graph by learning a more effective adaptive graph.

Remark: The objective function in Eq. (4) simultaneously seeks a semantic encoder through marginalized denoising strategy and latent semantic representation guided with adaptive graph reconstruction. In this way, our semantic encoder can benefit the artificial data augmentation to span the visual feature space of seen classes. Also, the adaptive graph reconstruction scheme could assist learning more effective latent semantic representation assuming the give semantics are not enough to describe the visual features. Two strategies tend to trigger each other to learn the semantic encoder with better generalization ability to unseen classes.

3.4. Optimization

It is straightforward to observe that three variables W, S and Z in Eq. (4) are not able to be jointly optimized. To deal with the issue, we first convert it into the augmented Lagrangian function via involving an extra variable E defined as $E = S - H$ ($S \geq 0$):

$$\begin{aligned} \mathcal{J} = &\mathbb{E}\Big(\|W\tilde{X} - Z\|_{\mathrm{F}}^2\Big) + \alpha\|Z - AS\|_{\mathrm{F}}^2 + \beta\|E\|_1 \\ &+ \mu\Big(\|\mathbf{1}_n^\top S - \mathbf{1}_n^\top\|_2^2 + \|S - H - E\|_{\mathrm{F}}^2\Big), \end{aligned} \tag{5}$$

To fight of the constraint $\mathbf{1}_n^\top S = \mathbf{1}_n^\top$ efficiently, we relax the constraint through incorporating a penalty term $\mu\|\mathbf{1}_n^\top S - \mathbf{1}_n^\top\|_2^2$ into Eq. (5) and μ is a positive parameter. Since the optimization of Eq. (5) is non-smooth and non-convex, and thus, we design an efficient solver to Eq. (5) with respect to W, S, Z and E, respectively.

Learning Semantic Encoder W: Given Z, the objective function w.r.t. W reduces to:

$$W = \underset{W^\top W = \mathrm{I}_m}{\arg\min}\ \mathbb{E}\Big(\|W\tilde{X} - Z\|_{\mathrm{F}}^2\Big), \tag{6}$$

where $\mathbb{E}(\tilde{X})$ can be calculated by following [18]. To fight off the non-convex problem in Eq. (6) due to the orthogonal constraint $W^\top W = \mathrm{I}_m$, we explore a gradient descent optimization [29]. In general, we first calculate the gradient of $\mathcal{J}$ w.r.t W as

$$\frac{\partial \mathcal{J}}{\partial W} = 2W\mathbb{E}\big(\tilde{X}\tilde{X}^\top\big) - 2Z\mathbb{E}(\tilde{X}^\top),$$

where $\mathbb{E}\big(\tilde{X}\tilde{X}^\top\big)$ and $\mathbb{E}(\tilde{X}^\top)$ can be calculated by following [18]. After that, we calculate the skew-symmetric matrix and update W until we reach the Armijo-Wolfe conditions [27].

Learning Adaptive Graph S: Given Z, E, we relax the non-negative constraint and rewrite the objective function w.r.t. S as:

$$\mathcal{J} = \|\bar{Z} - \bar{A}S\|_{\mathrm{F}}^2 + \mathrm{tr}(\Gamma S^\top), \tag{7}$$

where $\bar{Z} = [\sqrt{\alpha}Z, \sqrt{\mu}\mathbf{1}_n, \sqrt{\mu}(H + E)]$ and $\bar{A} = [\sqrt{\alpha}A, \sqrt{\mu}\mathbf{1}_n, \sqrt{\mu}\mathrm{I}_n]$. For constraint $S \geq 0$, we introduce the Lagrange multiplier Γ, which is an extra variable. Fortunately, we can mitigate the optimization of Γ through the following deduction. To be specific, we obtain the partial derivative of $\mathcal{J}$ over S and set it to zero as:

$$\frac{\partial \mathcal{J}}{\partial S} = 2\bar{A}^\top(\bar{A}S - \bar{Z}) + \Gamma = 0.$$

Algorithm 1 Solving the problem in Eq. (5)

Input: X, A, H, α, β

Initialization: $\mu_m = 10^6, \mu_0 = 10^{-1}, \rho = 1.3, \tau = 0$.

while not converged **do**

1. Update Z via Eq. (10) with others fixed.
2. Update S, E via Eqs. (8) and (9) with others fixed.
3. Update W via Eq. (6) with others fixed.
4. Update penalty $\mu_{\tau+1} = \min(\rho\mu_\tau, \mu_m)$.
5. Check the convergence condition $|\mathcal{J}_{\tau+1} - \mathcal{J}_\tau| < 10^{-3}$.
6. $\tau = \tau + 1$.

end while

output: W, S, E, Z.

Through the KKT condition $\Gamma \odot S = 0$ ($\odot$ means the Hadamard product), we can obtain the following formulation:

$$\Big[2\bar{A}^\top(\bar{A}S - \bar{Z}) + \Gamma\Big] \odot S = 0.$$

Following [37], we obtain the updating rule for S:

$$S = S \odot \sqrt{\frac{\bar{A}^\top \bar{A} S}{\bar{A}^\top \bar{Z}}}, \tag{8}$$

where we mitigate the optimization of Γ.

After we optimize S, E could be further updated with the following l_1-optimization problem:

$$\begin{aligned} E = & \arg\min_E \beta\|E\|_1 + \mu\|S - H - E\|_{\mathrm{F}}^2 \\ = & \ \mathrm{sign}\Big(S - H\Big)\max\Big(|S - H| - \frac{\beta}{2\mu}, 0\Big). \end{aligned} \tag{9}$$

Learning Latent Semantics Z: Given W, S, we can update Z by minimizing Eq. (10) w.r.t Z:

$$\begin{aligned} Z = & \arg\min_Z \mathbb{E}\Big(\|W\tilde{X} - Z\|_{\mathrm{F}}^2\Big) + \alpha\|Z - AS\|_{\mathrm{F}}^2 \\ = & \ \frac{1}{\alpha+1}\Big(W\mathbb{E}(\tilde{X}) + \alpha AS\Big). \end{aligned} \tag{10}$$

For better clarity, we present the optimization details in **Algorithm 1**, where we list the initialization of some variables. To ensure a good convergence, we initialize W with the mapping between X and A. Other variables are initialized with random matrices for simplicity. α and β are two hyper parameters, which would be selected based on the validation set.

In ZSL tasks, there are different cases to do evaluation. For zero-shot recognition, we are to predict their class label given any reference visual data. Considering a test data x_t^i, we could first calculate its predicted semantic embedding with Wx_t^i using semantic encoder, then compare with the ground-truth semantic representation A_t with C_t classes (C_t would cover both seen classes C and unseen classes C_u). For zero-shot annotation, we just exploit the predicted the semantics to search its attributes through several largest values. For zero-shot retrieval, we would adopt the given semantics a_t to search the most similar visual samples over predicted semantics WX_t.

Table 1. Statistics of four ZSL benchmarks.

Dataset	aP&aY	AwA2	CUB	SUN
#Training Categories	20	40	150	645
#Test Categories	12	10	50	72
#Samples	15,339	37,322	11,788	14,340
#Semantics	64	85	312	102
#Training Samples	5,932	23,527	7,057	10,320
#Test Seen Samples	1,483	5,882	1,764	2,580
#Test Unseen Samples	7,924	7,913	2,967	1,440

4. Experiment

In this part, we conduct experiments on four ZSL benchmarks, by comparing our proposed approach with state-of-the-art ZSL from conventional and generalized ZSL tasks.

4.1. Dataset & Experimental Setting

Four zero-shot learning benchmarks are evaluated in our experiments including SUN attribute dataset[1], Animals with Attributes 2 (AwA2)[2], Caltech-UCSD Birds 2011 (CUB)[3] and aPascal-aYahoo (aP&aY)[4]. Their statistics are provided in Table 1. All these benchmarks are served with annotated attributes.

Due to some unseen test categories in the original splits for those four benchmarks belong to part of ImageNet [24], Xian et al. recently proposed a new split protocol [32], targeting at a true zero-shot evaluation. In our experiments, we strictly follow the split protocol and adopt the 2048-D ResNet-101 features for all four benchmarks [32]. Moreover, we utilize the continuous attributes for better ZSL performance.

For our model with the k-nn graph, we adopt $k = 10$ as default across various ZSL tasks simply. The trade-off parameters are chosen from the range $[10^{-2}, 10^2]$ according to the evaluation performance on the labeled samples from the seen categories in the validation set. Later on, we directly utilize the selected parameters to conduct evaluation on the original seen and unseen classes. Because different initializations would result in different optimal solutions for our proposed model, and we run five times of our model and report the average results per task.

Baselines: The comparisons with the state-of-the-art include DAP/IAP [16], CONSE [19], CMT [25], SSE [35],

[1] http://cs.brown.edu/~gmpatter/sunattributes.html

[2] https://cvml.ist.ac.at/AwA2/

[3] http://www.vision.caltech.edu/visipedia/CUB-200-2011.html

[4] http://vision.cs.uiuc.edu/attributes/

Table 2. Conventional zero-shot recognition in terms of top-1 accuracy (%) on SUN, CUB, AWA2 and aP&aY benchmarks with ResNet visual features.

Method	SUN	CUB	AwA2	aP&aY
DAP [16]	39.9	40.0	46.1	33.8
IAP [16]	19.4	24.0	35.9	36.6
CONSE [19]	38.8	34.3	44.5	26.9
CMT [25]	39.9	34.6	37.9	28.0
SSE [35]	51.5	43.9	61.0	34.0
LATEM [30]	55.3	49.3	55.8	35.2
ALE [1]	58.1	54.9	62.5	39.7
DEVISE [10]	56.5	52.0	59.7	39.8
SJE [2]	53.7	53.9	61.9	32.9
ESZSL [23]	54.5	53.9	58.6	38.3
SYNC [5]	56.3	55.6	46.6	23.9
SAE [12]	40.3	33.3	54.1	8.3
PSR [3]	61.4	63.8	56.0	38.4
ZSKL [34]	61.7	51.7	**70.5**	45.3
Ours	**62.8**	**64.2**	67.8	**46.2**

LATEM [30], ALE [1], DEVISE [10], SJE [2], ESZSL [23], SYNC [5], SAE [12], PSR [3], and ZSKL [34]. The last two are the most recently proposed ZSL algorithms. PSR also aims to explore the relation structure by mining the most similar and dissimilar pairs, thus could learn a more discriminative metric. ZSKL attempts to learn a non-linear mapping across the visual feature and attribute spaces by exploring kernel functions. Note that results are directly copied from other published papers, i.e., [32, 3, 14], since we explore the exactly same protocol and the same set of data. Moreover, the approaches encompass a wide range in zero-shot learning area.

Evaluation Metric: Top-1 accuracy is widely-used to measure single-label classification accuracy. That is the prediction is correct for the assignment class label equals to the ground-truth one. In zero-shot learning, top-1 accuracy per-class is more valued, since high performance is encouraged in both densely and sparsely populated categories. Hence, we average the accurate predictions independently for each category before dividing their cumulative sum, w.r.t the number of categories [32].

For generalized zero-shot learning (GZSL) scenario, the search space during the evaluation stage is not only restricted to the unseen classes (U), but also consists of the seen classes (S). Thus, the harmonic mean[5] is more popular to measure the GZSL performance by calculating the average per-class top-1 accuracy on training and test categories [32]. This strategy is able to flag up those ZSL models overfitting to either seen or unseen classes.

[5] https://en.wikipedia.org/wiki/Harmonic_mean

4.2. Conventional Zero-shot Recognition

This section reports the comparison results (Table 2) on conventional zero-shot recognition in terms of top-1 accuracy. From Table 2, we witness that our proposed algorithm is able to obtain better performance by comparing with others. This verifies that our approach learns a more effective visual-semantic relation from seen data for unseen data analysis. The obtained improvements are very consistent according to the complexity of visual images of each benchmark, which e can observe from the well-known complicated CUB dataset. On other hand, our model still performs very well on SUN benchmark, which contains more classes and relatively fewer training instances per class. For AwA2, only class-wise attributes are provided, thus it is challenging for our model to recover the missing attributes by exploring the relation across different instances and categories.

Compared with PSR and ZSKL, which explore non-linear neural networks or kernel functions to link visual and semantics, our model also preserves such non-linear property. Since we attempt to learn a latent semantic representation, it builds a bridge to link the visual features and provided semantics. Especially, we utilize an adaptive graph to reconstruct the latent semantics. All these provide more flexibility to the learned generic encoder and thus is able to improve the generalizability on unseen classes.

Furthermore, qualitative results are reported for our designed model. We aim to list what kinds of visual information our algorithm is able to capture only given the semantic representation for the unseen categories. Figure 2 reports 10 out of 50 unseen categories in CUB dataset, in which we show top-3 accurately-retrieved samples (middle row in red) while the top-3 misclassified samples (last row in blue) into each unseen class. Observing from the top images, the proposed model reasonably discovers discriminative visual information for each unseen category only using its semantic representation. We further notice that the misclassified visual images have much different visual appearances to that of assigned class. Hence, it is hard to recognize them, even for humans.

4.3. Generalized Zero-shot Recognition

In a more general application, we are not sure if the test image belongs to the seen categories or totally unseen categories, which is more interesting from a practical point of view. In this sense, a lot of research efforts focus on the generalized zero-shot challenge, in which the test set are built on both seen and unseen category data.

Table 3 reports the generalized ZSL performance of all comparisons, where $U \rightarrow U+S$ and $S \rightarrow U+S$ represent two types of GZSL that evaluate if learned unseen/seen models are confused to each other. H denotes the harmonic mean. From Table 3, we can easily notice that generalized ZSL results are significantly lower than conventional ZSL

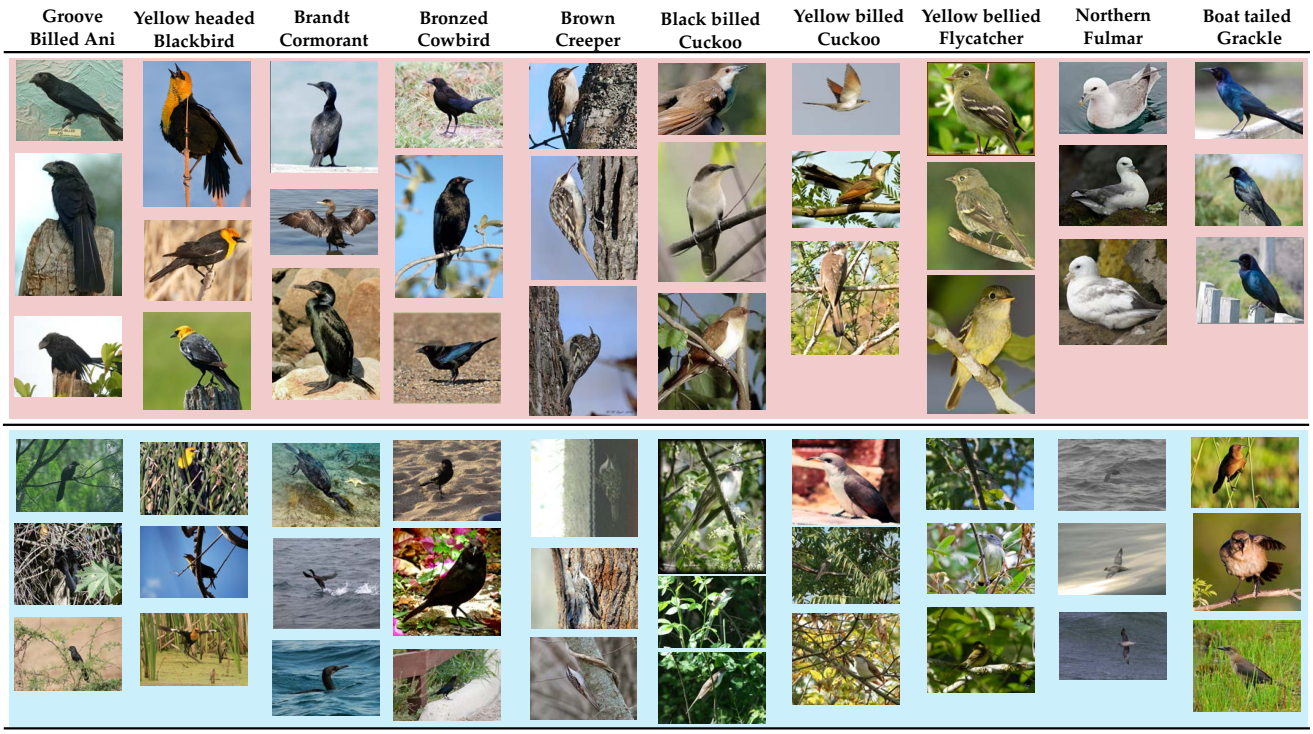


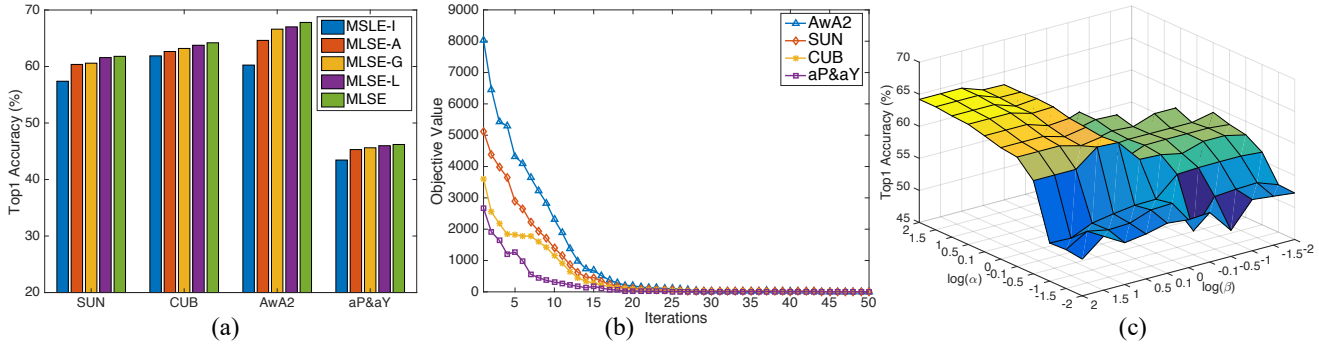
Figure 2. Qualitative evaluation of our proposed model on CUB benchmark, in which 10 unseen category labels are listed on the top. Then, we report the top-3 samples assigned to each category in the middle. Finally, the last row shows the top-3 misclassified instances.

Table 3. Generalized ZSL recognition performance (%) across four benchmarks.

Method	SUN			CUB			AwA2			aP&aY		
	U→S+U	S→S+U	H	U→S+U	S→S+U	H	U→S+U	S→S+U	H	U→S+U	S→S+U	H
DAP [16]	4.2	25.1	7.2	1.7	67.9	3.3	0.0	84.7	0.0	4.8	78.3	9.0
IAP [16]	1.0	37.8	1.8	0.2	72.8	0.4	0.9	87.6	1.8	5.7	65.6	10.4
CONSE [19]	6.8	39.9	11.6	1.6	72.2	3.1	0.5	90.6	1.0	0.0	91.2	0.0
CMT [25]	8.1	21.8	11.8	7.2	49.8	12.6	0.5	90.0	1.0	1.4	85.2	2.8
SSE [35]	2.1	36.4	4.0	8.5	46.9	14.4	8.1	82.5	14.8	0.2	78.9	0.4
LATEM [30]	14.7	28.8	19.5	15.2	57.3	24.0	11.5	77.3	20.0	0.1	73.0	0.2
ALE [1]	21.8	33.1	26.3	23.7	62.8	34.4	14.0	81.8	23.9	4.6	73.7	8.7
DEVISE [10]	16.9	27.4	20.9	23.8	53.0	32.8	17.1	74.7	27.8	4.9	76.9	9.2
SJE [2]	14.7	30.5	19.8	23.5	59.2	33.6	8.0	73.9	14.4	3.7	55.7	6.9
ESZSL [23]	11.0	27.9	15.8	12.6	63.8	21.0	5.9	77.8	11.0	2.4	70.1	4.6
SYNC [5]	7.9	43.3	13.4	11.5	70.9	19.8	10.0	90.5	18.0	7.4	66.3	13.3
SAE[12]	8.8	18.0	11.8	7.8	54.0	13.6	1.1	82.2	2.2	0.4	80.9	0.9
PSR[3]	20.8	37.2	26.7	20.7	73.8	32.3	24.6	54.3	33.9	13.5	51.4	21.4
ZSKL [34]	20.1	31.4	24.5	21.6	52.8	30.6	18.9	82.7	30.8	10.5	76.2	18.5
Ours	20.7	36.4	26.4	22.3	71.6	34.0	23.8	83.2	37.0	12.7	74.3	21.7

ones. This results from the fact that seen categories are included in the search space, that play as distractors for the unseen samples.

An interesting phenomenon is that compatibility learning algorithms, e.g., DEVISE, ALE and SJE, are able to obtain good performance on unseen classes. However, these approaches perform well on seen classes, since they seek independent attribute or object classifiers, e.g., DAP and CONSE. Compared with these methods, our proposed model also achieves very competitive results in each metric, especially in harmonic mean measurement. In terms of the harmonic mean measurement, our proposed approach performs the best on SUN, AwA2, and aP&aY datasets while the second best on CUB dataset, where ALE outperforms others). This also verifies the effectiveness of our proposed approach in generalized ZSL tasks.



(a) (b) (c)

Figure 3. (a) evaluation on different variants; (b) convergence curves of our proposed algorithm for four ZSL tasks; (c) parameter influence of α and β using CUB ZSL task.

4.4. Empirical Analysis

First of all, we evaluate on several variants of our proposed MLSE to dive deeper to the efficacy of each component. 1). MLSE-L denotes we use F-norm to replace the l_1-norm in Eq. (4); 2). MLSE-G means Eq. (3); 3). MLSE-A represents we use a pre-defined graph G instead in Eq. (3) (i.e., $\beta = 0$); 4). MLSE-I is the version that we set S as the identity matrix. Then, we conduct experiments on four benchmarks and report the comparison results in Figure 3(a), where we notice that the performance drops significant when we directly enforce latent semantics Z to be close to the given A. The performance increases a lot with a graph reconstruction format, which denotes the graph reconstruction is capable of compensating the attributes across various samples and categories. Moreover, the adaptive graph could contribute to enhancing the performance over different ZSL tasks, which means the adaptive graph is able to automatically capture the relationship across the latent semantics and the given semantics. Finally, we also witness the improvements with a sparse l_1-sparse regularizer.

Secondly, we show our model's convergence from experimental side empirically. The convergence curves of four benchmarks on our proposed algorithm are presented in Figure 3 (b), where we observe that our model has a good convergence after several iterations, especially after 40 iterations. The experimental results show our model can converge well.

Thirdly, we testify the parameter influence in terms of recognition performance to evaluate the two novel regularizers. We jointly evaluate α and β on CUB tasks with ResNet features. From Figure 3 (c), we notice that the recognition performance would increase with the values of α and β becoming larger, which indicate both parameters play an important role in our semantic encoder.

Finally, we visualize 10 unseen AwA2 categories with their learned latent semantics Z using ResNet features as the input. To be specific, we explore t-SNE[6] to embed the learned latent semantics of the unseen data points to a 2-D plane in Figure 4. From the results, we notice there are most classes are well separate, while some samples from unseen classes are overlapped. This indicates our model is valid in generalizing to unseen classes.

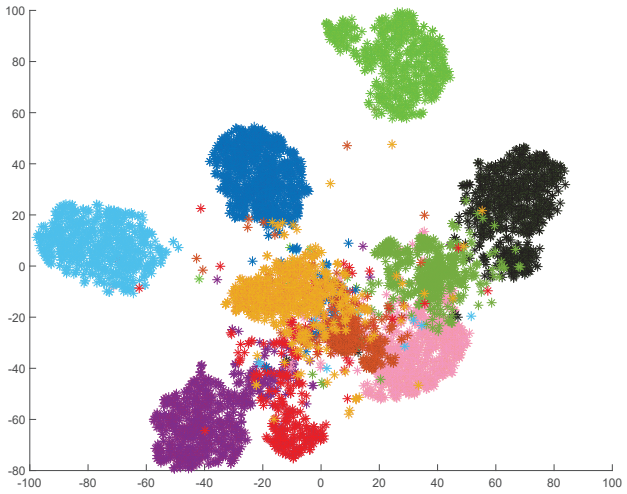


Figure 4. Visualization of 10 unseen class data points from AwA2 with the learned semantics Z. The same color denotes the same category data points.

5. Conclusion

In this paper, we developed a novel zero-shot learning algorithm through learning adaptive latent semantic representation. To be specific, we presented an effective knowledge transfer model by jointly seeking a generic semantic encoder and learning latent semantic representation. To augment the visual space of seen classes, we exploited a marginalized denoising strategy to cover the unseen classes. Furthermore, we sought an adaptive reconstruction coefficient to learn the latent semantic representation by capturing more intrinsic information from the given semantics. Conventional and generalized ZSL evaluations on four ZSL benchmarks were testified to demonstrate the effectiveness of our proposed marginalzied semantic encoder.

[6] https://lvdmaaten.github.io/tsne/

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *TPAMI*, 38(7):1425–1438, 2016. 2, 6, 7

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015. 2, 6, 7

[3] Y. Annadani and S. Biswas. Preserving semantic relations for zero-shot learning. In *CVPR*, pages 7603–7612, 2018. 1, 2, 6, 7

[4] M. Bucher, S. Herbin, and F. Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *ECCV*, pages 730–746. Springer, 2016. 1

[5] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336, 2016. 2, 6, 7

[6] S. Changpinyo, W.-L. Chao, and F. Sha. Predicting visual exemplars of unseen classes for zero-shot learning. In *ICCV*, pages 3476–3485, 2017. 1

[7] L. Chen, H. Zhang, J. Xiao, W. Liu, and S.-F. Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding network. In *CVPR*, volume 2, 2018. 1, 2

[8] Z. Ding, M. Shao, and Y. Fu. Low-rank embedded ensemble semantic dictionary for zero-shot learning. In *CVPR*, pages 2050–2058, 2017. 1

[9] Z. Ding, M. Shao, and Y. Fu. Generative zero-shot learning via low-rank embedded semantic dictionary. *TPAMI*, 2018. 1, 2

[10] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NeurIPS*, pages 2121–2129, 2013. 6, 7

[11] H. Jiang, R. Wang, S. Shan, Y. Yang, and X. Chen. Learning discriminative latent attributes for zero-shot classification. In *CVPR*, pages 4223–4232, 2017. 1, 2

[12] E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, pages 4447–4456, 2017. 1, 2, 6, 7

[13] S. Kolouri, M. Rostami, Y. Owechko, and K. Kim. Joint dictionaries for zero-shot learning. In *AAAI*, pages 3431–3439, 2018. 1

[14] V. Kumar Verma, G. Arora, A. Mishra, and P. Rai. Generalized zero-shot learning via synthesized examples. In *CVPR*, pages 4281–4289, 2018. 1, 6

[15] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009. 2

[16] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 36(3):453–465, 2014. 5, 6, 7

[17] Y. Liu, Q. Gao, J. Li, J. Han, and L. Shao. Zero shot learning via low-rank embedded semantic autoencoder. In *IJCAI*, pages 2490–2496, 2018. 1, 2

[18] L. Maaten, M. Chen, S. Tyree, and K. Weinberger. Learning with marginalized corrupted features. In *ICML*, pages 410–418, 2013. 3, 4

[19] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. 2014. 5, 6, 7

[20] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, pages 503–510. IEEE, 2011. 2

[21] P. Peng, Y. Tian, T. Xiang, Y. Wang, and T. Huang. Joint learning of semantic and latent attributes. In *ECCV*, pages 336–353. Springer, 2016. 2

[22] G.-J. Qi, W. Liu, C. Aggarwal, and T. Huang. Joint intermodal and intramodal label transfers for extremely rare or unseen classes. *TPAMI*, 39(7):1360–1373, 2017. 1

[23] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 2152–2161, 2015. 1, 2, 6, 7

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 5

[25] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NeurIPS*, pages 935–943, 2013. 5, 6, 7

[26] J. Song, C. Shen, J. Lei, A.-X. Zeng, K. Ou, D. Tao, and M. Song. Selective zero-shot classification with augmented attributes. In *ECCV*, pages 468–483, 2018. 1

[27] W. Sun and Y.-X. Yuan. *Optimization theory and methods: nonlinear programming*, volume 1. Springer Science & Business Media, 2006. 4

[28] X. Wang, Y. Ye, and A. Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, pages 6857–6866, 2018. 1

[29] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013. 4

[30] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77, 2016. 6, 7

[31] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018. 1, 2

[32] Y. Xian, B. Schiele, and Z. Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, pages 3077–3086, 2017. 5, 6

[33] X. Xu, F. Shen, Y. Yang, D. Zhang, H. T. Shen, and J. Song. Matrix tri-factorization with manifold regularizations for zero-shot learning. In *CVPR*, pages 3798–3807, 2017. 1

[34] H. Zhang and P. Koniusz. Zero-shot kernel learning. In *CVPR*, pages 7670–7679, 2018. 1, 2, 6, 7

[35] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, pages 4166–4174, 2015. 5, 6, 7

[36] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, pages 6034–6042, 2016. 2

[37] H. Zhao, Z. Ding, and Y. Fu. Multi-view clustering via deep matrix factorization. In *AAAI*, pages 2921–2927, 2017. 5

[38] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, pages 1004–1013, 2018. 1, 2